Capstone Project report

Introduction

Background:

Mumbai and Delhi are the two most important metro cities in India. There has always been a comparison in terms of quality of life, jobs, education, entertainment and recreational facilities that these cities have to offer to its residents. This project elaborates on a data science project that attempts to analyse the neighbourhoods in each of these two cities and tries to understand what is popular in them and what they have to offer to someone who is contemplating to make a choice about living in either of the metro cities.

The deciding factor for most would be on how lively, supportive, vibrant and unique each of the cities can be when compared to each other. The business problem in this study assumes that people who would be interested in this study are those who would like to create a projection of potential life and activities in these metro city neighbourhoods if the subject moves to live in one of them. The decision to choose one over the other would depend on popular venues in the neighbourhoods in each of these metro cities.

Problem Statement:

Analyse the data regarding venues in Delhi and Mumbai. Cluster these neighbourhoods and draw conclusions as to what the

Data Acquisition & Cleaning

Data Source

For this study, we will need data about neighbourhoods in each of these metro cities. The data published by the government on postal codes for all India would serve us well for this study. We will specifically download the CSV provided under https://data.gov.in/resources/all-india-pincode-directory-contact-details-along-latitude-and-longitude.

In this study, we will download the CSV, read it into a pandas Data frame and curate it to remove the data related to all other cities, towns, and places which are not Mumbai or Delhi, since we are only interested in comparing these two biggest metro cities in India.

We shall then clean up the unnecessary columns in the CSV, which is not relevant or useful for our current study. Post office names (office name) will be used as the neighbourhood names in each of the regions such as Mumbai or Delhi.

Neighbourhood names with the same Pin code will be combined as a single row.

Foursquare API will be used to find the longitude and latitude of each of the neighbourhoods in both Mumbai and Delhi. This will form the dataset we will use for this study.

Combining these sources of data we will cluster the neighbourhoods to group them and understand them in depth.

Data Cleaning

Data downloaded from various sources were combined into a dataframe. The dataframe was checked for missing or NaN values. Unnecessary columns were removed from the data frame.